GENERAL TIME REVERSIBLE DISTANCES WITH UNEQUAL RATES ACROSS SITES

by

PETER J. WADDELL[*] and M.A. STEEL[#]

No. 143 May, 1996

[*] School of Biological Sciences, Massey University, Palmerston North, New Zealand (Corresponding Author)

[#] Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch, New Zealand

Abstract

A series of new results, with proofs, useful to the study of DNA sequences using Markov models of substitution are presented. General time-reversible distances can be extended to accommodate any fixed distribution of rates across sites by replacing the logarithmic function of a matrix, with the inverse of a moment generating function. Estimators are presented assuming a gamma distribution, the inverse Gaussian distribution, or a mixture of either of these with invariant sites. Also considered are the different ways invariant sites may be removed, and how these differences may affect estimated distances. The variance of these new distances is approximated via the delta method. It is also shown how to predict the divergence expected for a pair of sequences given a rate matrix and distribution of rates across sites, so allowing iterated ML estimates of distances under these models. A simple test of whether a rate matrix is time reversible is also presented, and this makes the identification of such models simple. These new methods are used to estimate the divergence time of humans and chimps from mtDNA sequence data. These analyses support suggestions that the human lineage has an enhanced transition rate relative to other hominoids. These studies also show that transversion distances differ substantially from the overall distances which are dominated by transitions. Transversions alone apparently suggest a very recent divergence time for humans versus chimps and / or a very old (>16myr) divergence time for humans versus orangutans. This work illustrates graphically ways to interpret the reliability of distance-based transformations, using the 'corrected' transition to transversion ratio returned for pairs of sequences which are successively more diverged.

Terms and abbreviations:

c	the sequence length
CSR	constant site removal (from the data)
δ	a transformed distance estimate
$\mathbf{F}$	a matrix of proportions of aligned paired nucleotides
$\mathbf{F}^{\#}$	matrix $\mathbf{F}$ symmetrised
i.r.	assuming all site are evolving at an identical intrinsic rate
p_{inv}	the proportion of invariant (invariable) sites
ti/tv	transition to transversion ratio

Introduction

Failure to allow for the unequal substitution rate at different sites in two aligned sequences can lead to serious underestimates of the true distance between them (Golding 1983). Furthermore, this underestimation becomes progressively worse the larger the true distance, which in turn compromises the additivity necessary for transformed distance phylogenetic methods to be guaranteed consistent (Felsenstein 1982; Felsenstein 1984; Felsenstein 1993). If this error becomes serious enough, parallelisms and convergences due to multiple substitutions at a site (which occur predominantly between long 'edges' of a tree) can outweigh parsimony 'informative' characters (Felsenstein 1978; Hendy and Penny 1989). This effect is often termed 'long edges attract', as such edges may be spuriously joined together by tree reconstruction methods (including distance methods), even when all other aspects of the model are correct (e.g. Hasegawa and Fujiwara 1993; Lewis and Gaut 1995; Waddell 1995; Chang 1996; Lockhart et al. 1996). Failure to account for unequal base compositions in the sequence also leads to a progressive underestimate of the 'true' distance (e.g. Tamura 1992), with similar effects expected upon tree selection.

The general time reversible distance is the most general transformation that can be applied to a pair of DNA sequences which aims to return the expected average number of substitutions per site. This was first described by Lanave et al. (1984), and in a different form by Tavaré (1986) and Rodríguez (1990). Since then it has been pointed out by Zharkikh (1994) that the Lanave et al. distance is based on the time reversible model, and not the general twelve parameter model they had claimed (while Lewis and Swofford unpublished, show their algebraic identity). Importantly, nearly all of the currently used distance estimates (including those of Tamura 1992; Tamura and Nei 1994) are special cases (restrictions) of the general time reversible distance (Zharkikh 1994; Swofford et al. 1996).

A general time reversible distance assumes a general time reversible model of evolution, which is a model where the probability (or likelihood) of the data is independent of the placement of the root on the tree (Felsenstein 1981, Adachi and Hasegawa 1994, Yang 1994). With the exception of some special matrices (the Kimura 3 ST being the most general, e.g. Evans and Speed 1993) this further implies the relative rates of all substitutions remains constant across the tree, and the root base composition is in equilibrium. This in turn implies the frequency of all states in the model (e.g. the four nucleotides A, C, G, and T) remain at the same frequencies; that is they (and the model) are said to be stationary. Given this model, it is possible to make estimates of the rates of all types of substitution using just pairs of sequences (Tavaré 1986). An extended introduction to these distances, and their relation to the more general 12 parameter models of DNA evolution (e.g. Barry and Hartigan 1987, Yang 1994) is given in the next section (see also Zharkikh 1994).

A variety of specific distances have been modified to take unequal substitution rates across sites into account. These include the Jukes-Cantor (1 parameter) and Kimura 2 parameter distances (Golding 1983; Olsen 1987; Jin and Nei 1990), a specific six parameter distance (Tamura and Nei 1993). In addition, a variety of methods to calculate likelihoods of the data under similar conditions have been used (Hasegawa et al. 1995; Churchill et al. 1992; Reeves

1992; Sidow et al. 1992; Steel et al. 1993; Yang 1993; Felsenstein and Churchill 1996; Waddell and Penny 1996). In this paper we show that for general time reversible distances, the necessary steps to accommodate a specified distribution of rates across sites are simple and can be made without the need to separate sites into rate classes. An important case of unequal rates across sites is the existence of some sites which are incapable of changing due to biological constraints. These invariant (invariable) sites lead to distortions of estimated distances (Shoemaker and Fitch 1989), and in some cases inconsistency of tree selection (Waddell 1995; Chang 1996; Lockhart et al. 1996; Waddell et al. 1996). Here we consider how time reversible distances may be modified to take these sites into account, especially when the base composition of these sites does not reflect that of the variable sites (Waddell 1995).

A primary motivation for this work was to have distances to both infer trees and more accurately estimate the edge lengths on trees i.e. a weighted trees. Weighted trees are critical for inferring the divergence times of many taxa (e.g. Hillis et al. 1996, Waddell and Penny 1996). We use an example from Waddell and Penny (based on 5kb of hominoid mtDNA sequences from Horai et al. 1992) to illustrate the new methods, and infer the divergence time of human versus chimp lineages.

Since 1994, following cooperative work with Dr David Swofford, time reversible distances allowing unequal rates across sites have been available in the computer package PAUP*. This program is the freely available beta test version of PAUP 4.0 (Swofford 1996) which also allows tree searching and bootstrapping using these and a wide variety of related distances (for further information contact swofford@onyx.si.edu).

Materials and Methods:

Time reversible distances: their form and assumptions

If all sites evolve at an identical rate (i.r.) the general time reversible distance can be written (Rodríguez et al. 1990) in the form,

$$\delta_{ij} = -\text{trace}(\Pi ln[\Pi^{-1}\mathbf{F}]) \qquad (1)$$

where δ_{ij} is the distance between sequences i and j measured as the expected number of substitutions per site (including multiple changes at a site), Π is a diagonal matrix of the nucleotide base composition of the sequences, $\mathbf{F}$ is the divergence matrix of sequences i and j, and ln is the matrix logarithm function. The divergence matrix is just the expected proportion of times one state is aligned next to another state in the two sequences (Fig. 1). The logarithm of a matrix $\mathbf{X}$ is defined as $ln(\mathbf{X}) = -\sum_{n=1}^{\infty} \frac{(\mathbf{I}-\mathbf{X})^n}{n}$, where $\mathbf{I}$ is the identity matrix, and provided this limit exists.

Under a time reversible process of evolution, all $\mathbf{F}$ matrices are symmetric in expectation (a proof is given in Appendix 1a). When dealing with finite samples, Π and $\mathbf{F}$ are replaced with their sample estimates (these are denoted by $\hat{\Pi}$ and $\hat{\mathbf{F}}$). Furthermore, $\hat{\mathbf{F}}$ is then replaced by $\mathbf{F}^{\#}$, a symmetrised form of $\hat{\mathbf{F}}$. This is done to reduce sampling errors, and $\mathbf{F}^{\#}$ is shown to be a ML

estimator of **F** under the model (see Appendix 1b). This result is convenient, since a useful way to evaluate the matrix logarithm function of the matrix $\Pi^{-1}\mathbf{F}^{\#}$ (if defined) is via diagonalisation (see Fig. 1). The symmetry of both $\mathbf{F}^{\#}$ and Π implies that their product is always diagonalisable and will have real eigenvalues (e.g. Keilson 1979).

These distances are consistent (i.e. become exactly correct as sequence lengths go to infinity), and so additive in expectation on a tree, provided all sites have the same rate of substitution (i.r. or identical rates) and the process of evolution is time reversible across all paths in the tree. The most general form of the rate matrix, **R**, then has nine parameters and can be written as $\Pi^{-1}\mathbf{S}$, where **S** is a symmetric matrix of relative rates, and Π is the diagonal matrix of the stationary base compositions of the nucleotide states (Tavaré 1986; a proof is given in Appendix 7). An equivalent parameterisation of **R** is $\mathbf{S}\Pi$, for a different but still symmetric rate matrix **S** (Tavaré 1986; Zharkikh 1994; with a proof appearing at the end of appendix 7). Thus Π has 3 free parameters (since nucleotide proportions must sum to 1), while **S** has up to 6 (since rows of a rate matrix sum to zero), making a total of 9 free parameters in the model.

A special case where equation (1) is also exact, but the model may not be strictly time reversible, is when the base composition is equal frequency [0.25, 0.25, 0.25, 0.25] and stationary throughout the tree (Rodríguez et al. 1990). To meet this requirement the **R** matrix must have both its rows and columns summing to zero. This gives rise to a total of 3 more constraints than a 12 parameter rate matrix, allowing up to 9 free parameters. However, not just any such matrix forms a time reversible model; a counter-example is the matrix

$$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array}\begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} -6 & 2 & 2 & 2 \\ 3 & -12 & 4 & 5 \\ 2 & 8 & -11 & 1 \\ 1 & 2 & 5 & -8 \end{bmatrix} \end{array}$$

(a test of reversibility is described later). Note, however, that the additional assumption of a molecular clock made by Lanave et al. (1984) and Rodríguez et al. (1990) in order for equation (1) to be exact is not necessary (e.g. Tavaré 1986; Barry and Hartigan 1987).

An example of calculating the general time reversible distance, is given in figure 1. It considers the divergence matrix between the human and chimp sequences of Horai et al. (1992) (as edited in Waddell and Penny (1996) with the removal of all sites with insertions or deletions). Evidence of non-stationarity of this data was sought using an exhaustive series of pairwise tests of base compositions using the X^2 statistic, as implemented in PAUP* (Swofford 1996). The overall result was non-significant (i.e. no evidence of non-stationarity), despite this test being expected to reject stationarity too often by ignoring the strong positive correlations amongst sequences caused by their common history (phylogeny). The test was also repeated after removing all constant columns from the data. This is a precaution since leaving invariant sites in will make the test less likely to reject non-stationarity. Again the result was not significant, despite the removal of all constant sites (and not just the minority likely to be invariant) being expected to bias the test towards rejection of stationarity.

$$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array}\begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} 1415 & 8 & 55 & 2 \\ 4 & 1371 & 1 & 144 \\ 73 & 0 & 578 & 0 \\ 3 & 117 & 1 & 1126 \end{bmatrix} \end{array} \quad \begin{bmatrix} 0.2889 & 0.0012 & 0.0131 & 0.0005 \\ 0.0012 & 0.2799 & 0.0001 & 0.0266 \\ 0.0131 & 0.0001 & 0.1180 & 0.0001 \\ 0.00005 & 0.0266 & 0.0001 & 0.2299 \end{bmatrix} \quad \begin{bmatrix} 0.3037 & 0 & 0 & 0 \\ 0 & 0.3079 & 0 & 0 \\ 0 & 0 & 0.1313 & 0 \\ 0 & 0 & 0 & 0.2571 \end{bmatrix}$$

$c\hat{\mathbf{F}}$ $\qquad$ $\mathbf{F}^{\#} = (\hat{\mathbf{F}} + \hat{\mathbf{F}}^{t})/2$ $\qquad$ $\hat{\Pi}$ $(\hat{\Pi}_{ii}$ = row sum of $\mathbf{F}^{\#})$

$$\begin{bmatrix} 0.9513 & 0.0040 & 0.0430 & 0.0017 \\ 0.0040 & 0.9092 & 0.0003 & 0.0865 \\ 0.0995 & 0.0008 & 0.8989 & 0.0008 \\ 0.0030 & 0.1036 & 0.0004 & 0.8940 \end{bmatrix} \quad \begin{bmatrix} 0.4088 & 0.5473 & -0.5000 & 0.0138 \\ 0.0020 & -0.4236 & -0.5000 & -0.6449 \\ -0.9134 & 0.5770 & -0.5000 & -0.0159 \\ -0.0164 & -0.4337 & -0.5000 & 0.7640 \end{bmatrix} \quad \begin{bmatrix} 0.8546 & 0 & 0 & 0 \\ 0 & 0.9922 & 0 & 0 \\ 0 & 0 & 1.0000 & 0 \\ 0 & 0 & 0 & 0.8066 \end{bmatrix}$$

$\hat{\mathbf{P}} = \hat{\Pi}\mathbf{F}^{\#}$ $\qquad$ Ω = right eigenvectors of $\hat{\mathbf{P}}$ $\qquad$ Ψ $(\psi_{ii}$ = eigenvalues of $\hat{\mathbf{P}})$

$$\begin{bmatrix} 0.7728 & 0.0038 & -0.7502 & -0.0264 \\ 0.6975 & -0.5473 & 0.3179 & -0.4681 \\ -0.6074 & -0.6158 & -0.2626 & 0.5143 \\ 0.0151 & 0.7137 & -0.0075 & 0.7661 \end{bmatrix} \quad \begin{bmatrix} -0.1571 \\ -0.0079 \\ 0.0000 \\ -0.2150 \end{bmatrix} \quad \begin{bmatrix} -0.0524 & 0.0042 & 0.0466 & 0.0016 \\ 0.0042 & -0.1008 & 0.0002 & 0.0963 \\ 0.1078 & 0.0006 & -0.1091 & 0.0007 \\ 0.00019 & 0.1154 & 0.0004 & -0.1176 \end{bmatrix}$$

Ω^{-1} $\qquad$ $\ln(\psi_{ii})$ $\qquad$ $\hat{\mathbf{R}}\tau = \Omega\ln(\Psi)\Omega^{-1}$

$$\begin{bmatrix} -0.0159 & 0.0013 & 0.0142 & 0.0005 \\ 0.0013 & -0.0310 & 0.0001 & 0.0297 \\ 0.0142 & 0.0001 & -0.0143 & 0.0001 \\ 0.0005 & 0.0293 & 0.0001 & -0.0302 \end{bmatrix} \quad \begin{bmatrix} -78.0 & 6.3 & 69.3 & 2.4 \\ 6.3 & -152.0 & 0.4 & 145.3 \\ 69.3 & 0.4 & -70.2 & 0.5 \\ 2.4 & 145.3 & 0.5 & -148.2 \end{bmatrix}$$

$\hat{\Pi}\hat{\mathbf{R}}\tau$ (subs. of each type per site) $\qquad$ $c\hat{\Pi}\hat{\mathbf{R}}\tau$ (estimated totals)

δ_{hc} is the expected number of substitutions per site between human and chimp sequences, which gives rise to the estimate $\hat{\delta}_{hc}$ = -trace($\hat{\Pi}\hat{\mathbf{R}}\tau$) = - (-0.0159 + -0.0310 + -0.0143 + -0.0302) = 0.09154 (0.09152 with full precision).

Fig. 1. The steps in calculating the time reversible distance (equation (1)). The observed divergence matrix $c\hat{\mathbf{F}}$ (where c is the sequence length) is for the comparison of human and chimp mtDNA sequences (Horai et al. 1992). Starting with the observed matrix of aligned paired-nucleotide frequencies ($c\hat{\mathbf{F}}$) we estimate $\mathbf{R}\tau$ and other quantities. Entries in $\hat{\mathbf{R}}\tau$ are inferred relative rates, whereas entries in $\hat{\Pi}\hat{\mathbf{R}}\tau$ are estimated numbers of each type of substitution divided by the sequence length. The observed (Hamming) distance from $\hat{\mathbf{F}}$ is $\sum_{i \neq j} F_{ij}$ = (8 + 55 + ... + 1) / 4898 = 0.0833, whereas the distance for the data corrected under the i.r. time reversible model is (6.3 + 69.3 + ... + 0.5) / 4898 = 0.0915. The matrix $c\hat{\Pi}\hat{\mathbf{R}}\tau$ (which is analogous to the $\hat{\mathbf{F}}$ matrix with corrections for multiple hits) shows the estimated number of transversions almost unchanged. In contrast, the number of multiple hits is estimated as [(69.3 + 69.3)/(55 +73) - 1] × 100% = 8.3% amongst the A ↔ G transitions, or [(145.3 + 145.3)/(117 + 144) - 1] × 100% = 11.3% amongst the more numerous C ↔ T transitions. This in turn has increased the overall transition to transversion ratio from 20.47 for the observed data to 22.50 for the i.r. time reversible model corrected data, an increase of 9.9%.

It is useful to test the expectation that $c\mathbf{F}$ is symmetric and so consider whether the data are showing a distinctly non-time reversible character, as suggested by Tavaré (1986). Using either a X^2 or G^2 test statistic is reasonable (Read and Cressie 1988). The X^2 test statistic is $\sum_{i\neq j}\frac{(c\hat{F}_{ij}-cF^{\#}_{ij})^2}{cF^{\#}_{ij}}$, while the G^2 test uses $\sum_{i\neq j}c\hat{F}_{ij}\ln\left(\frac{\hat{F}_{ij}}{F^{\#}_{ij}}\right)$. Both test statistics asymptotically have a χ^2 distribution with degrees of freedom (d.f.) equal to number of entries $i \neq j$ (12) minus the number of estimates made in $\mathbf{F}^{\#}$ (6), which leaves 5 d.f. For the comparison of human and chimp sequences, the X^2 and G^2 values are 8.86 (P = 0.11) and 4.84 (P = 0.56) respectively (or 6.86 (P = 0.14) and 3.45 (P = 0.49) when grouping the cells with expected values of less than 1, leaving 4 d.f.). None of these are significant. However, a problem with this test is its lack of power when a molecular clock is likely, since this too implies that $\mathbf{F}$ is symmetric (see proof in Appendix 1c). This type of test also has the non-independence and invariant site problems mentioned earlier for the pairwise base composition tests.

Results:

The time reversible distance with a distribution of rates across sites

Distances estimated under stationary time reversible models (with up to 9 parameters in their transition matrices) can be extended to allow for unequal rates across sites using the same general approach used in Steel et al. (1993) and Waddell et al. (1996) (for the Hadamard conjugation). As explained below, this extension allows correction for a variety of site rate distributions including the commonly used Γ and lognormal. Our new distance formula estimating the expected number of substitutions per site is,

$$\delta_{ij} = -\text{trace}(\Pi M^{-1}[\Pi^{-1}\mathbf{F}]) \qquad (2)$$

where M^{-1} is the inverse of the moment generating function of the distribution of rates across sites (defined below and see a proof in Appendix 2). The application of M^{-1} to $\Pi^{-1}\mathbf{F}$, (here taken as matrix $\mathbf{Z}$) is defined as,

$$M^{-1}(\mathbf{Z}) = \Omega M^{-1}[\Psi]\Omega^{-1} \qquad (3)$$

where Ω is a matrix containing, as columns, the right eigenvectors of $\mathbf{Z}$ (i.e. $\mathbf{Z}\Omega = \Omega\mathbf{D}$), Ω^{-1} is its inverse, and function M^{-1} is applied componentwise to the diagonal entries of the diagonal matrix Ψ of the associated eigenvalues of Z. As with the time reversible i.r. model, we symmetrise $\hat{\mathbf{F}}$ to give $\mathbf{F}^{\#}$ when dealing with sampled data. It is possible to prove that any distance based on only the observed dissimilarity (e.g. that of Jukes-Cantor or Tajima-Nei; see Swofford et al. 1996) and assuming identical site rates will always underestimate the true distance if there is any site-to-site rate variation (Appendix 3).

The function M[x] is defined as the expectation, $M[x] = E[e^{\lambda_j x}]$, the moment generating function of the statistical distribution the λ_j (Table 2 of Waddell et al. 1996 gives some specific examples). Note that $M[x] \approx \frac{1}{c}\sum_{i=1}^{c} e^{\lambda_i x}$, the average value of the $e^{\lambda_i x}$ over the

sites (where c is the sequence length). Here, the argument of M will always be ≤ 0 (rather than positive as in most statistical applications). Consequently function M will always be defined in our applications, and lie in the range 0 to 1. M^{-1} denotes the left functional inverse (the standard inverse) of M (i.e. $M^{-1}[M[x]] = x$), which always exists since $M[x]$ is a monotone increasing function.

Due to sampling error when estimating **F** from a finite number of sites, the eigenvalues of **P** (= $\Pi^{-1}\mathbf{F}^{\#}$) (which are expected to lie in the range [0, 1]), may be negative, making M^{-1} undefined (basically the distance appears too large, or infinite given the expectations of the model). This is a commonly encountered problem with all model based distance transformations, which may also be caused by non-stationarity of base composition (Waddell 1995). In such cases, a useful rule of thumb when estimating phylogenetic relationships, is to set these undefined distances to twice the value of the largest defined distance from the distance matrix of species being compared. This is justified since the largest distance (path) on an additive tree can never be more than twice the size of the second largest value. Given more information about the tree, it may also be possible to refine the expected range for inapplicable distance estimates. In real applications we do not know the function M exactly for any given sequence, so its form is estimated with a method that compares more than 2 sequences at a time, as discussed later.

This general approach also provides a quick way of calculating the transition matrix, **P**, along any edge or path moving down a tree when rates at sites vary and **R** is given (e.g. when modeling sequence evolution). Let τ be equal to the total expected number of substitutions on an edge or along a path, while **R** is scaled so that the positive entries of $\Pi\mathbf{R}$ sum to 1, then

$$\mathbf{P} = M[\mathbf{R}\tau] \tag{4}$$

This last result allows us to quickly calculate the divergence matrix (**F**) under any continuous time Markov process. As with equations (2) and (3), it is assumed sites evolve independently. A proof of the last equation is given in Appendix 4. It is useful to note that if **R** defines a time reversible process it can always be diagonalised and has real eigenvalues (Keilson 1979, section 3.2). For convenience we will label the eigenvalues of $\mathbf{R}\tau$ as entries ξ_{ii} of the diagonal matrix Ξ (e.g. Fig. 1 and 2).

The modification of equation (1) to allow for unequal site rates, equation (2), is achieved by the replacing the natural logarithm function by the function M^{-1}. Waddell (1995) and Waddell and et al. (1996) list the moment generating functions, plus their inverses, for commonly used distributions which may approximately describe site rates. Specifically, these moment generating functions are for standardised distributions (e.g. Stuart and Ord 1987, p. 192) where the mean of the underlying distribution has been set to 1 (i.e. $E_\lambda[\lambda] = 1$) so that inferred distances are recovered as the expected number of substitutions per site and not some other multiple of this number (e.g. Golding 1983; Jin and Nei 1989; Steel et al. 1993). Two distributions are particularly useful because they both have closed forms for both M and M^{-1}. The first of these is for the much used gamma (Γ) distribution (e.g. Golding 1983, Jin and Nei 1990, Steel et al. 1993), where $M[x] = ((k - x) / k)^{-k}$, while $M^{-1}[x] = k(1- x^{-1/k})$. Here k is the shape parameter of the Γ distribution. When $k \rightarrow \infty$, this distribution tends to the delta distribution (i.e. identical rates),

and M tends to the ln function. When k decreases the distribution assumes a skewed normal shape: at $k = 1$ it has progressed to the shape of the exponential distribution, and for $k < 1$ the distribution becomes ever more 'L' shaped (and site rates more uneven, e.g. Golding 1983, Jin and Nei 1990, Swofford et al. 1996). The ratio of the standard deviation to the mean of the λ_j (the coefficient of variation, or c.v.) is $k^{-0.5}$ (the standard deviation, since the mean is fixed to 1).

The second distribution, from Waddell et al. (1996), is the inverse Gaussian distribution which is shaped like the lognormal distribution (introduced with genetic distances by Olsen 1987). For the inverse Gaussian, $M[x] = \exp(d\{1- [1- (2x /d)]^{0.5}\})$, while $M^{-1}[x] = 0.5d(1- \{1- (ln[x] /d)\}^2)$. Here d is the shape parameter, and the coefficient of variation for site rates is $d^{0.5}$. Here, as $d \rightarrow \infty$, M tends to the natural logarithm function. As d decreases below 1, the rates across sites follow a highly skewed lognormal-like distribution. Apart from the notable shape difference of the Γ versus the inverse Gaussian distribution at rates near zero, the inverse Gaussian distribution also tends to have a 'flatter' tail than the Γ, inferring more of the most rapidly evolving sites (e.g. the sites evolving more than 40 times the mean rate). Distributions such as the lognormal, F and Weibull show this second feature even more prominently. Distributions with flat tails, will often infer more multiple hits (i.e. the assumed set of very rapidly evolving sites) and accordingly will often also infer higher ti/tv ratios (since multiple transitions are most likely 'hidden').

Constant site removal and invariant sites distributions

Not accounting for invariable sites leads to distortions of estimated distances (Shoemaker and Fitch 1989) and sometimes inconsistency of tree selection (Waddell 1995; Lockhart et al. 1996). If we could identify these sites, they could be edited out. In reality their identity is usually uncertain, however it is still possible to accommodate a proportion of invariant sites, and this can always be done at the **F** level given estimates of their overall proportion and the relative frequencies of the invariant bases (Waddell 1995). Here the definition $\mathbf{F}_{var} = (\mathbf{F} - p_{inv}\Pi_{inv}) / (1-p_{inv})$ is used, where $\mathbf{F}_{var}$ is the **F** matrix of just the variable sites, p_{inv} is the inferred proportion of invariant (unable to change) sites, and Π_{inv} is the diagonal matrix of the base composition of the invariant sites.

There are a number of ways to specify Π_{inv}, the base composition of the invariant sites: (1) The simplest is that the invariant sites will have base composition equal in the 4 bases, i.e. $\pi_{inv} = [0.25, 0.25, 0.25, 0.25]$. (2) Often a more likely possibility is that the invariant sites will reflect the base composition of the sequences as a whole. This can be done in two ways. We can take $\pi_{inv} = \pi$ (for a particular **F** matrix). However, if we are confident of the applicability of the earlier model, then π should be stationary across all sequences, so π can be used for all comparisons as the average base composition across all sequences. (3) A more robust approximation is that the base composition of the sites which are unvaried (constant in all sequences) better reflects π_{inv} (tested on the Horai data in Fig. 2 caption). Again this modification can be done by estimating π_{inv} for each pair of sequences, or preferably with regards to accuracy and sampling error, estimated as an average from the constant sites constant across all sequences. (4) The vector π_{inv} could also be supplied from an ML program that estimated base composition

of all sequences, or which separately optimised the base composition of the varied and unvaried sites. This last option is most computationally intensive, but if the model holds relatively well, is expected to be the most preferable for statistical accuracy. The first three ways of making these modifications are available in PAUP* (Swofford 1996), while the fourth is in preparation.

An interesting feature is that the more the base composition of the varied sites and the unvaried sites differ, often the more pronounced the amount of correction made when π_{inv} is inaccurately estimated (i.e. over-estimation of distances if the model were to hold exactly). To avoid this we tend to prefer estimating π_{inv} as the base composition of the sites which are constant across all sequences (via method 3 or 4).

When using these corrections as a general modification to give robustness to unequal rates across sites, the term a 'Constant Site Removal' (CSR) modification, seems appropriate. Three CSR methods are used in later examples: CSR(F) (a form of method 2), where the base composition of the invariant sites are estimated separately as π for each pairwise **F** matrix, CSR(all sites) (again a variant of method 2) where π_{inv} is estimated as the unweighted average across all sequences, and CSR(cons.) (method 3) where π_{inv} is estimated from the sites constant across all taxa .

The various CSR modifications are most easily made before forming $\mathbf{F}_{var}$ from the observed 'paired nucleotide' counts. However, with the situation of π_{inv} estimated as π for each pair of sequences, the modification to the transformation of the eigenvalues can be made by replacing $ln[x]$ with $ln[(x - p_{inv}) / (1 - p_{inv})]$, or more generally $M^{-1}[(x - p_{inv}) / (1 - p_{inv})]$ (Waddell et al. 1996). Thus, either way, it is straight forward to allow for mixed variable / invariable site distributions (e.g. Gu et al. 1995; Waddell 1995; Waddell et al. 1996; Waddell and Penny 1996).

ML estimators and iterated ML distances

Simulations and analytic calculations with equation (2) suggest it yields an ML estimate of the true distance given just $\hat{\mathbf{F}}$ and a distribution of rates across sites. By ML estimate, we mean the distance which will minimise the G^2 statistic between **F** and $\hat{\mathbf{F}}$ when all entries in **R** (i.e. both components of $\mathbf{R} = \Pi^{-1}\mathbf{S}$) are simultaneously optimised. ML estimators often have the desirable property that as c becomes large, they have the minimum possible sampling variance of all estimators for that model. Consistent with this, under models known to have ML distance estimators (e.g. those of Jukes and Cantor 1969, Kimura 1980 and 1981, as shown in Saitou 1990) equation (5) returns identical variance estimates to the delta method variances of these estimators. In our simulations and bootstrap analyses, equation (2) often has less variance than distance estimators which are known not to be ML estimators. These include the 3P distance of Tamura (1992) and the six parameter distance of Tamura and Nei (1993) (see Zharkikh 1994). In this sense it is often a better distance to use for estimating evolutionary trees than these other estimators, especially when distances become larger and / or base composition unequal (but stationary).

It is also possible to use equation (4) to predict what the expected divergence matrix would be given the observed base composition, $\hat{\Pi}$, an estimate of **R** normalised so that all off-diagonal entries in $\Pi\mathbf{R}$ sum to 1, and the function M. In this case the expected divergence matrix is just $\mathbf{F}_{exp} = \Pi\mathbf{P}_{exp} = \Pi M[\mathbf{R}\tau]$. This allows the likelihood of the observed pairwise data, $c\hat{\mathbf{F}}$, matrix to be calculated then optimised by likelihood in the same way that Felsenstein (1993) does for the more specific i.r. Kimura 2P (1980) and Felsenstein (1984) distances (in the program DNADIST). This approach is expected to give lower sampling errors under the model if the necessary parameters (**R**, plus the distribution of rates across sites and so also function M) can be estimated by a statistically efficient method such as ML applied to a set of sequences. This approach will be discussed in more detail elsewhere, and is also available in PAUP* (Swofford 1996).

A computational example

Application of equation (2) to the mtDNA sequences of Horai et al. (1992) allows correction for an unequal distribution of rates across sites, a very high transition to transversion ratio, and a skewed base composition (see caption to Fig. 2) of mammalian mtDNA. To illustrate some consequences of using equation (2) and different distributions of rates across sites, the same human-chimp comparison as in figure 1 is used. Three different distributions of rates across sites are used, with shape parameters for each distribution being calculated by maximum likelihood (ML) tree estimation on sequences allowing for a distribution of rates across sites. This was done using the ML methods and models of Waddell and Penny (1996) (for a proof of the site likelihood calculations see Steel et al. 1993 or Yang 1993). Using the generalised Kimura 3ST model is both computationally convenient and unlikely to result in overestimates of the spread of site rates (Waddell 1995).

With the edited Horai et al. (1992) data, the ML sequence based method estimated the shape parameter of the inverse Gaussian distribution as $d = 0.213$. The optimal fit of data to model, measured by the likelihood ratio, G^2 (Ritland and Clegg 1987, Stuart and Ord 1991, p. 1160) was 334.8 (Waddell unpublished). For the Γ distribution, k was estimated to be 0.351 and G^2 303.2, while an invariant sites / i.r. distribution yielded $p_{inv} = 0.592$ and the best fit of $G^2 =$ 279.4 was achieved (Waddell and Penny 1996). Allowing a mixture of invariant sites with either a Γ or an inverse Gaussian distribution did not further improve fit. In all cases the proportion of invariant sites went to its optima, and the continuous distribution took on a shape parameter to mimic an i.r. distribution. Figure 2 and Table 1 show results of using equation (2) to infer aspects of mtDNA evolution between humans and chimps.

M^{-1} eigenvalues	Inferred rate matrix per nucleotide	Overall inferred substitutions
$M^{-1}[\psi_{ii}]$	$\hat{\mathbf{R}}\tau$	$c\hat{\Pi}\hat{\mathbf{R}}\tau$

(a) M^{-1} for inverse Gaussian with $d = 0.213$, $\delta = 0.13274$.

$$\begin{bmatrix}-0.2151\\-0.0080\\0.0000\\-0.3235\end{bmatrix}\quad\begin{bmatrix}-0.0707 & 0.0053 & 0.0643 & 0.0012\\0.0052 & -0.1507 & -0.0002 & 0.1457\\0.1487 & -0.0004 & 0.1489 & -0.0006\\0.0014 & 0.1745 & 0.0003 & -0.1762\end{bmatrix}\quad\begin{bmatrix}-105.2 & 7.8 & 95.6 & 1.8\\7.8 & -227.3 & -0.3 & 219.7\\95.6 & -0.3 & -95.7 & 0.4\\1.8 & 219.7 & 0.4 & -221.9\end{bmatrix}$$

(b) M^{-1} for Γ with $k = 0.213$, $\delta = 0.12205$.

$$\begin{bmatrix}-0.1982\\-0.0080\\0.0000\\-0.2966\end{bmatrix}\quad\begin{bmatrix}-0.0654 & 0.0050 & 0.0591 & 0.0013\\0.0049 & -0.1384 & -0.0001 & 0.1335\\0.1368 & -0.0002 & -0.1373 & 0.0007\\0.0015 & 0.1598 & 0.0004 & -0.1617\end{bmatrix}\quad\begin{bmatrix}-97.3 & 7.4 & 87.9 & 1.9\\7.4 & -208.6 & -0.1 & 201.3\\87.9 & -0.1 & -88.3 & 0.4\\1.9 & 201.3 & 0.4 & -203.6\end{bmatrix}$$

(c) M^{-1} for with $p_{inv} = 0.592$, Π_{inv} estimated from $\mathbf{F}_{hc}$, $\delta = 0.26713$ (0.10899).

$$\begin{bmatrix}-0.4406\\-0.0194\\0.0000\\-0.6427\end{bmatrix}\quad\begin{bmatrix}-0.1461 & 0.0115 & 0.1312 & 0.0034\\0.0113 & -0.3003 & 0.0001 & 0.2888\\0.3034 & 0.0003 & -0.3056 & 0.0018\\0.0040 & 0.3458 & 0.0009 & -0.3508\end{bmatrix}\quad\begin{bmatrix}-88.6 & 7.0 & 79.6 & 2.1\\7.0 & -184.8 & 0.1 & 177.7\\79.6 & 0.1 & -80.2 & 0.5\\2.1 & 177.7 & 0.5 & -180.3\end{bmatrix}$$

(d) M^{-1} for with $p_{inv} = 0.592$, Π_{inv} estimated from constant sites, $\delta = 0.26635$ (0.10867).

$$\begin{bmatrix}-0.5772\\-0.0202\\0.0000\\-0.5575\end{bmatrix}\quad\begin{bmatrix}-0.1687 & 0.0125 & 0.1523 & 0.0040\\0.0098 & -0.2543 & 0.0001 & 0.2445\\0.3993 & 0.0002 & -0.4018 & 0.0023\\0.0042 & 0.3251 & 0.0009 & -0.3301\end{bmatrix}\quad\begin{bmatrix}-93.2 & 6.9 & 84.1 & 2.2\\6.9 & -179.4 & 0.0 & 172.4\\84.1 & 0.0 & -84.6 & 0.5\\2.2 & 172.4 & 0.5 & -175.1\end{bmatrix}$$

Fig. 2. The effect of different forms of the distribution of rates across sites on the transformed distances.

(A) An inverse Gaussian (with shape parameter $d = 0.213$), (B) Γ (with shape $k = 3$), (C) CSR(F) and (D) CSR(cons) are invariant sites / variable sites distributions. Averaging over all 6 hominoid mtDNA sequences in Horai et al. (1992) gives $\pi = [0.30, 0.31, 0.13, 0.26]$, whereas just the unvaried sites have base composition $\pi^c = [0.32, 0.28, 0.15, 0.25]$, which is significantly different (by a X^2 statistic test). For both invariant sites models, the estimated rate matrix is for just the variable sites. Likewise, δ, measured over just the variable sites is shown first, then δ averaged over all sites (i.e. multiplied by $(1-p_{inv})$) is given in brackets.

Table 1. Distances and ti/tv estimates with different distributions of rates across sites

rate distribution	δ_{hc}	increase over i.r.	ti/tv	increase over observed	ratio tr(AG) / tr(CT)	% multiple hits in tv's
i.r.	0.0915	-	22.50	9.9%	0.477	0.4%
inverse Gaussian	0.1327	45.0%	32.34	58.0%	0.435	2.6%
Γ	0.1221	33.4%	29.90	46.1%	0.437	1.8%
CSR(F)	0.1090*	19.1%	26.77	30.8%	0.448	1.2%
CSR(cons.)	0.1087*	18.7%	26.68	30.3%	0.488	1.2%

As figure 2 and Table 1 show, after taking a distribution of rates across sites into account, the estimated distance between these sequences increases substantially over that estimated assuming identical rates (i.r.) at all sites. Importantly, the size of this correction is dependent upon the assumed distribution, here being minimal with the invariant sites models, and largest with the inverse Gaussian distribution. As the distance between sequences increases, the effect is even more profound (see later). In contrast, very few multiple hits amongst the transversions are predicted to have occurred. A partial solution to the uncertainty of which distribution to use, is to ignore those models which have a substantially lower likelihood than the optimal model. Here, this would suggest ignoring the Γ, inverse Gaussian and i.r. estimates. This solution is only partial since our optimal model may still not be the true model, which could have a distinct distribution of rates, and could therefore infer quite different distances again. Additionally, the model being used to measure likelihoods may be critically deficient in some way, and consequently not be accurately measuring the rank of models attributable to the distribution of rates over sites. Thirdly, we implicitly assume the relative rates of sites are fixed, but they are unlikely to be (e.g. a covarion model like that of Fitch and Markowitz 1970, seems more likely, Waddell and Penny 1996) so all fixed site rate estimates could be severely mislead at larger distances.

Different forms of site rate distribution also give distinct ratios of transitions to transversions. In this case the largest ratio for the human-chimp comparison was with the inverse Gaussian distribution, consistent with this distribution having the longest tail.

Two distinct invariant sites models are evaluated in figure 2. While both give similar estimates of δ, the matrix $c\Pi\mathbf{R}\tau$ shows this is partly coincidence. This second model is suggesting there is a smaller proportion of variable sites with A or G than there are with C or T. This in turn leads to greater correction of AG transitions and fewer for CT transitions, but coincidentally these two effects nearly cancel (they do not always, as we see later). We do not show a mixed invariant sites / continuous distribution (such as CSR with Γ), since for this data such a mixture did not improve the likelihood of the models considered (Waddell and Penny 1996). However, the properties of such corrections (e.g. inferred distance, or ti/tv ratio) tend to be an average of their component parts (Waddell 1995). Overall then, the type of distribution of rates across sites is an important parameter to gaining more precise estimates of absolute distances (and hence exact edge lengths on trees), as well as estimated ti/tv ratios. While here the invariant sites model fits the data better than either a Γ distribution or an inverse Gaussian distribution, this is not always the case (Waddell 1995).

A delta method approximation of the variance

The previous sections have dealt with estimating the expected distance between a pair of sequences, but the variances of these estimates are also important. It is possible to derive an approximate variance formula for all these time reversible distances based on the delta method, which is frequently used in statistics (e.g. Stuart and Ord 1987, p. 324). With long sequence lengths and / or small distances this formula returns values very close to the true variance. Barry and Hartigan (1987) derived a delta method approximation to the variance of the i.r. time

reversible distance. It is straight forward to extend this approach for equation (2) which allows unequal rates across sites (proven in Appendix 5). This gives the result,

$$Var[\hat{\delta}] \approx \frac{1}{c}\left[\sum_{k=1}^{4}\pi_k(R_{kk}-\sum_{i=1}^{4}\pi_i R_{ii})^2+\sum_{k=1}^{4}\pi_k\left\{\sum_{l=1}^{4}P_{kl}\left(G_{kl}-\sum_{j}P_{kj}G_{kj}\right)^2\right\}\right]+O(c^{-2}) \qquad (5)$$

where ($\mathbf{G}_{kl}$) are elements of the matrix, $\mathbf{G}=-\sum_{r=1}^{\infty}a_r\sum_{s=0}^{r-1}\mathbf{B}^s(\mathbf{B}^t)^{r-1-s}$, $\mathbf{B} = \mathbf{I} - \mathbf{P}$ (where $\mathbf{B}^t$ is the transpose of $\mathbf{B}$), while $\mathbf{P} = \mathbf{\Pi}^{-1}\mathbf{F}^{\#}$, $\mathbf{R} = M^{-1}[\mathbf{P}]$ (replaced by their estimates when working from a finite sample) and c is the sequence length. The term which changes given different assumed distributions of rates across sites is a_r. This term is given by the coefficients of the Taylor series expansion of $M^{-1}[1\text{-}x] = -\sum_i a_i x^i$. For example, in the case of the i.r. distribution, a_r is the coefficient of x^r in $-ln(1\text{-}x)$, and so $a_r = 1/r$ (i.e $a_1 = 1$, $a_2 = 1/2$, $a_3 = 1/3$, ...). For the Γ distribution with shape parameter k the series becomes, $a_r = [(k+1)(2k+1)...((r-1)k+1)]/(r!k^{r-1})$, for example, with shape parameter $k = 0.351$, $a_1 = 1/(1 \times 0.351^0) = 1$; $a_2 = (0.351+1)/(2!(0.351^1)) = 1.92$; $a_3 = [(0.351+1)(0.702+1)]/(3!(0.351^2)) = 3.11$. As k goes to infinity, this series converges to that for the i.r. distribution, as expected.

For the inverse Gaussian distribution, $a_r = \frac{1}{r}+\frac{1}{2d}\sum_{m=1}^{r-1}\frac{1}{m(r-m)}$, as derived in Appendix 6. As an example, with shape parameter $d = 0.213$, this gives $a_1 = 1 + 1/0.426 \times \sum_{m=1}^{1-1}\frac{1}{m(1-m)} = 1$ + 0 (since the summation does not take effect) = 1; $a_2 = 1/2 + 1/0.426 \times \sum_{m=1}^{2-1}\frac{1}{m(2-m)} = 1/2 +$ $1/0.426 \times (1) = 2.85$; $a_3 = 1/3 + 1/0.426 \times (1/2 + 1/2) = 2.68$, $a_4 = 1/4 + (1/3 + 1/4 + 1/3)/0.426 =$ $0.250 + 1.49 = 2.40$, etc. If we look at the terms for this distribution, there are two parts. The first part, i.e. 1/r, is the same as the standard i.r. log transform, while the second part can be thought of as extra uncertainly due to unequal rates. As $d \rightarrow \infty$, then this second term goes to zero, and the variance converges to that of the i.r. model, as expected.

With the CSR distances, the easiest ways to make the computation of this variance is to redefine $\mathbf{P}$, $\mathbf{R}$ and Π as $\mathbf{P}_{CSR}$, $\mathbf{R}_{CSR}$ and Π_{CSR} i.e. their values after the removal of constant sites. For the mixed invariant sites / Γ distribution (Gu et al. 1995; Waddell 1995; Waddell and Penny 1996; Waddell et al. 1996) or a mixed invariant sites / inverse Gaussian distribution (Waddell 1995; Waddell et al. 1996) do as for the CSR distances, except apply the appropriate power series for the term a_r in equation (5).

Applying this variance to the mtDNA data is informative. The major cost in calculating $Var[\hat{\delta}]$ is evaluating the matrix $\mathbf{G}$. The determination of when the summation used to calculate $\mathbf{G}$ can be truncated depends on the data and distribution. In the HC comparison the entries in the second part of the summation (the products of matrix $\mathbf{B}$) quickly gives a pattern where entries decrease by a factor of approximately 4 with each increase in index r. This, combined with a

quickly decreasing term for a_r (as in the i.r. distribution), makes **G** calculated with r up to 4 accurate to the third decimal place. However, with larger distance comparisons to the siamang, the summation involving products of matrix **B** yields terms decreasing more slowly (by approximately 1/2 for the first 10 terms). In addition, for a Γ distribution and $k < 1$, the term a_r steadily increases with r, thereby requiring many more terms to determine **G** accurately (11 in this instance, for similar accuracy). Failure to consider enough terms in estimating **G** results in equation (5) returning a smaller than expected value for the variance. The very first summation term of equation (5), $\sum_{k=1}^{4} \pi_k (R_{kk} - \sum_{i=1}^{4} \pi_i R_{ii})^2$, tends to be small, e.g. in the HC comparison, assuming the inverse Gaussian distribution, it had value 0.0018 versus 0.41 for the remaining terms.

The increase of the standard errors with unequal rates across sites can be substantial. With the i.r. time reversible distance the estimated HC distance has a standard error of 0.0048. Whereas, the CSR(F) distribution of site rates (p_{inv} = 0.592) gives a s.e. of 0.0066, a slight decrease in accuracy. Assuming the Γ (k = 0.351) increases the inferred standard error to 0.00837, while the inverse Gaussian gives 0.00915. In contrast, the simpler i.r. Kimura (1980) 2ST (s.e. = 0.0047) or Jukes-Cantor (1969) (s.e. = 0.0044) distances give only a slight reduction in stochastic error.

Clearly, an unequal distribution of rates across sites can substantially decrease the accuracy of distance estimation even with a distance as short as that from human to chimp. In most instances this can be expected to decrease the bootstrap support for trees in comparison to the use of i.r. distances, and so unequal rates across sites must be considered in evaluating the robustness of any phylogenetic tree (Waddell 1995). This theoretical prediction is becoming a serious concern with many real data sets, which estimated under i.r. assumptions look highly informative, but when considering unequal site rates are not giving significant support (e.g. Waddell 1995; Lockhart et al. 1996). Additionally, equation (5) returns a monotonically increasing variance the larger the coefficient of variation (s.d. divided by mean) for a distribution of rates across sites becomes. Thus, all other factors equal, a sequence with a lower coefficient of variation of rates across sites will return a more reliable distance estimate if the **F** matrix is otherwise identical.

Lastly, as is generally the case with delta method approximations of the variance of a distance, equation (5) assumes the form of distribution, shape parameter(s) and base composition of invariant sites are known. The variability of distances due to estimating these parameters is best taken into account in a bootstrapping procedure, which includes re-estimating the shape parameter(s) with each bootstrap sample.

Applying time-reversible transformations to mtDNA from apes and humans

Some applications and implications of the general time reversible model with distributions of rates across sites are illustrated with the six taxon 5kb hominoid mtDNA of Horai et al. (1992). The first application is to estimate the ti/tv ratio between pairs of taxa (Fig. 3). When unequal rates across sites are allowed for, the inferred ratio increases substantially over

either that observed or that estimated by i.r. time reversible distances. The increase is 20% for the chimp-pygmy chimp comparison (with an inverse Gaussian distribution), and this increases steadily with increasing observed distance, reaching on average about 100% in distances measured to the outgroup siamang (the maximum being 110% for the HS CSR(all) sites transformation).

The data in figure 3 show an anomalously high ti/tv ratio for the HC and HP comparisons, and also less strikingly for the HG comparison (taxa abbreviations as given in Fig. 3). This suggests an increased ti/tv ratio in the human lineage, which is all the more striking as it goes against the trend of decreasing ti/tv ratio with increasing distance. This observation is consistent with the claims of Adachi and Hasegawa (1996). As regards the overall trend of decreasing ti/tv ratio, we doubt this is really as strong as it appears, rather it is most likely an artifact of these estimators in relation to the true model of evolution. That this trend is so pronounced, yet so consistent amongst all the assumed distributions of rates across sites, suggests that in general it may be very difficult to obtain accurate estimates of the ti/tv transition rate across all sites when taxa are even moderately diverged. This may give an indication of how accurately a distance transformation is correcting for multiple hits amongst the transitions. The observed and i.r. transformations data appear to be most strongly underestimating the expected ti/tv substitution ratio. In contrast, the inverse Gaussian ratios generally tended to be larger than the Γ distance ti/tv ratios. We expect this is due to a "flat tails effect", where the inverse Gaussian predicts there to be more of the most rapidly evolving sites than does the Γ (see Waddell et al. 1996). A lognormal distribution of site rates is expected to show this effect even more markedly.

The methods of constant site removal are showing another trend described in Waddell et al. (1996). Here, as the removal of constant sites brings the "infinite distance" (i.e. one or more of the eigenvalues tend to zero) closer (i.e. approaches an asymptote), then total distance and relative rates and ratios can, for relatively small changes in d_{obs}, experience a great increase. The effect here is not dramatic, but it is expected to become more pronounced for 'deeper' comparisons e.g. human to monkey sequences.

There are ways to alleviate, but not eliminate this problem of a downward bias in estimating relative rates with some substitutions going very fast hence incurring many multiple hits. One is more judicious editing of the data, separating out the three codon positions, from structural RNA coding regions (and possibly splitting the later into partitions such as stems and loops). This would not only serve to homogenize the relative rates of sites in each class, but might also show up distinct differences in the ratio of ti/tv changes in different regions. Such heterogeneity of the ti/tv ratio between sites could be an explanation for the trend seen in figure 3a. Another possibility is that the shape parameters of the various distributions are underestimated. This does occur here as seen by ML estimates with other mechanisms of evolution (Waddell 1995), but does not alter the overall trend (data not shown). Elsewhere, we diagnose these sequences further and show that by considering just 4 fold degenerate sites, there is much less bias in estimated ti/tv ratios at larger distances. Our purpose here is just to illustrate

how substantial this effect can be amongst taxa diverged less than 20 million years, so implying that many inferences of ti/tv ratios at older times may be of dubious value.

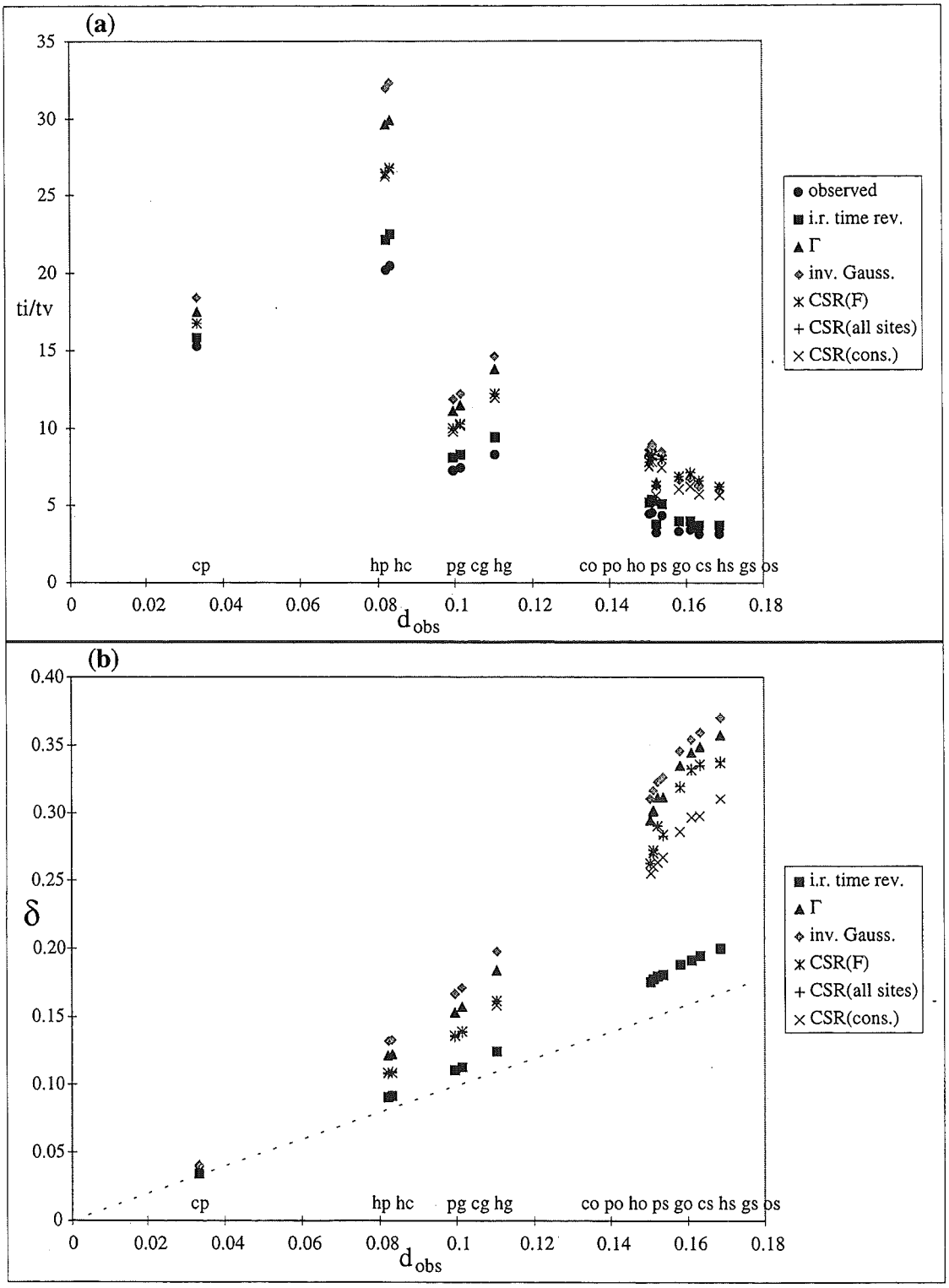

Fig. 3. (a) Transition versus transversions ratios and (b) total distance with rates across sites time reversible distances versus the observed (Hamming) distance (d_{obs}). The pair of sequences being compared is indicated in rank order by the symbols c (chimpanzee), p (pygmy chimp), h (human), g (gorilla), o (orangutan), and s (siamang). Using simpler distances, such as those of Jukes-Cantor or Kimura (2ST), yield even lower ti/tv ratios than the i.r. general time reversible used in (a). The dotted line in (b) shows the expected distance if there was no correction for multiple hits being made. The addition of unequal rates across sites has more than tripled the number of 'corrections' being made, relative to the standard i.r. time reversible distances.

While the general time reversible distances are essentially an ML method applied separately along all paths in the tree, a full ML for more than two species should be more robust in estimating the ti/tv ratio given branch points breaking up long paths (just as parsimony does). This is because the descendants of that node give more direct knowledge of ancestral states (e.g. Farris 1973). Unfortunately, most currently available ML programs assume a fixed ti/tv ratio across the tree and so would not indicate the apparently higher ti/tv ratio associated with humans in this data (unless subsets of taxa were analysed).

Figure 3b shows that all of these transformations assuming unequal rates across sites are making increasingly large corrections for multiple hits compared to the i.r. methods, as the observed distance becomes larger (as expected from the proof in appendix 3). There is also a noticeable difference between the different distributions, with the removal of constant sites making a lesser difference initially, but CSR(all sites) and CSR(F) showing signs of gaining on the Γ and inverse Gaussian models. This trend of CSR distances at some point becoming of larger magnitude than distances from other distributions is described and explained in Waddell et al. (1996). As mentioned earlier, this behaviour is related to the transformation approaching an asymptote, before becoming undefined (i.e. one or more of the eigenvalues of $\Pi^{-1}(\mathbf{F}-p_{inv}\Pi_{inv})$ go to zero).

The major impetus to developing these distances was to extend the methods for making divergence time inferences of humans and apes beyond those used in Waddell and Penny (1996). A simple way to infer a divergence time of a group (e.g. human-chimpanzee) is the ratio of a distance going through the node of interest (e.g. human-pygmy chimp), relative to a distance going through a node of more reliably known age (e.g. chimpanzee-orangutan). For the six taxa in this study, all such ratios are shown for all possible pairs of distances (Fig. 4a) estimated with the various distances. This figure shows the substantial, and closely agreeing, drop in divergence time estimates achieved by all the methods modeling a distribution of rates across sites. The dates fluctuate little with respect to the distance to orangutan used (due no doubt to a high correlation of paths through the tree), slightly more with respect to the use of chimpanzee or pygmy chimp sequence, but substantially when there is a choice of distance between human-orangutan versus chimp-orangutan. Clearly there can be expected to be even greater fluctuations with choice of species when estimating older divergence times. The ages on the right y-axis are made assuming a 16 million year old divergence of orangutans form African apes. This is a date preferred by biogeographic and fossil evidence as interpreted in Waddell and Penny (1996). The divergence dates in Waddell and Penny (1996), made with ML assuming both a Γ and invariant sites distribution of site rates (under a Kimura 3P model), fall between the i.r. time reversible dates, and the dates when rates across sites are allowed for (excluding the dates when the human-gorilla distance is used).

Given the appearance of a molecular clock for these data (Horai et al. 1992; Adachi and Hasegawa 1994; Waddell and Penny 1996), it is surprising that the divergence times estimated from just the transversion substitutions are so different (Fig. 4b). Since the model used to correct the transversion distance is making little difference to these times, they are not easily explained as systematically biased. Taken at face value they indicate either incredibly recent divergence

times for these taxa (especially humans from chimps) or a very ancient divergence of orangutan. Either interpretation runs into conflict with the fossil evidence. The nature of this apparent conflict is considered further elsewhere (Waddell in preparation).

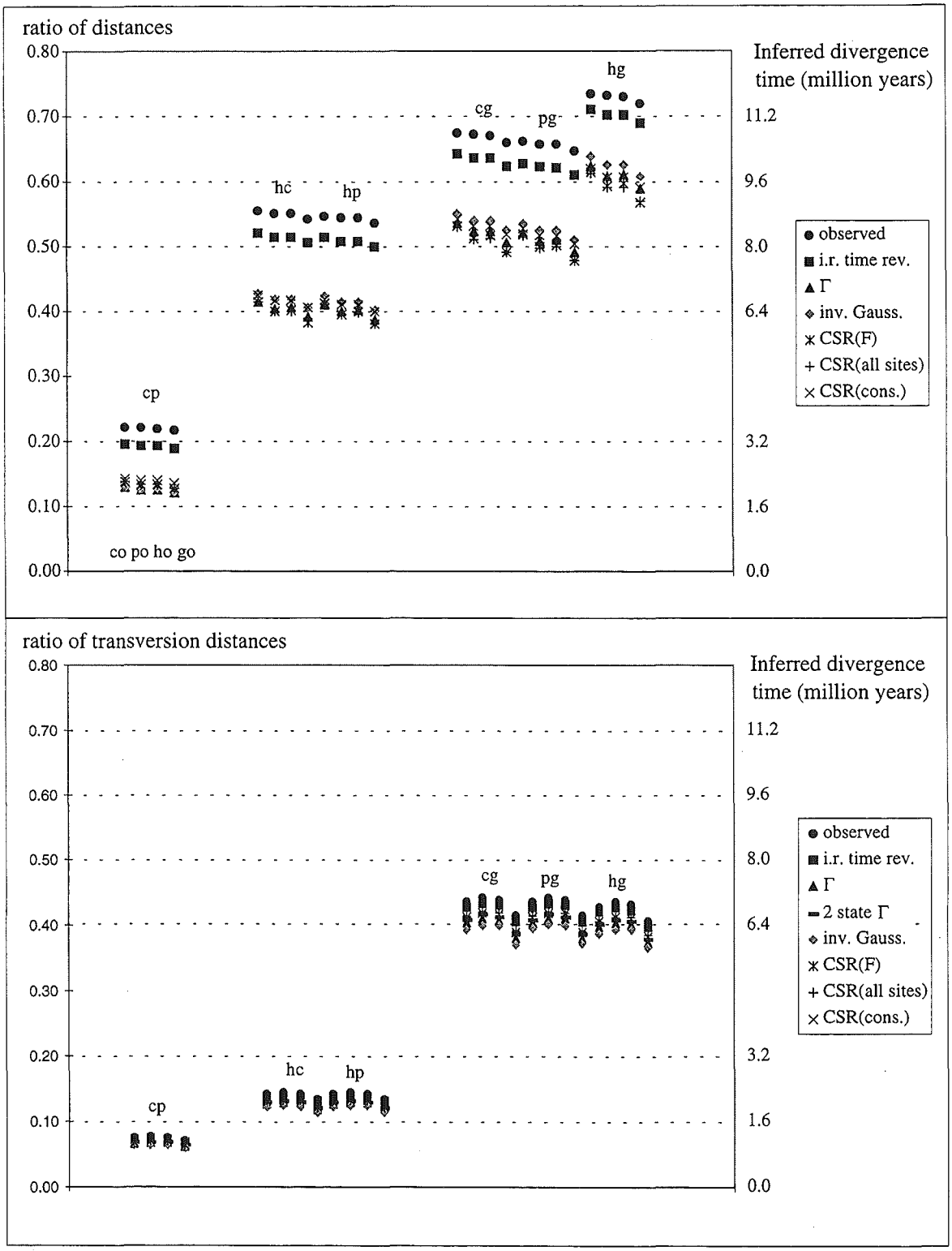

Fig. 4a Estimates of divergence times using time reversible distances counting all types of substitution and Fig. 4b just transversional changes. In Fig. 4a, all possible pairs of distances (for the more recent and the older divergence) are shown (x-axis is arbitrary). The distance used corresponding to the more recent divergence (e.g. cp) is shown above the block of 4 values, and the denominator for this comparison (e.g. the distance co, po, ho or go) is shown below the first set, with the same order in all instances. Fig. 4b follows the same pattern of pairs of distances, but uses only the transversional changes. The distance

transformations are the same as Fig. 4a, except for the mapping down to purines / pyrimidines (AG versus CT) Cavender distance is also shown. The divergence time is calibrated by assuming the orangutan - African apes split occurred 16 million years ago for the reasons outlined in Waddell and Penny (1996).

A quick test of the reversibility of a specified rate matrix

Often rate matrices are derived *de novo* without explicit reference to the stationary base composition, or whether the rate matrix can be written in the form of $\Pi^{-1}\mathbf{S}$ (or the equivalent form $\mathbf{S}\Pi$, e.g. Zharkikh 1994). However, a simple test is that rate matrix $\mathbf{R} = \begin{bmatrix} * & A & B & C \\ D & * & E & F \\ G & H & * & I \\ J & K & L & * \end{bmatrix}$ (all non-diagonal elements > 0, and row sums = 0), defines a time reversible model if, and only if, the following three conditions hold:

(E1) AGE = BDH, (6)

(E2) AJF = CDK, (7)

(E3) EKI = FHL. (8)

A proof of this test is given in Appendix 7. It is these three constraints that drop the number of free parameters in **R** from 12 down to a maximum of 9 for time reversible models. Relatively few examples of non-time reversible models have been proposed to study molecular evolution. They include: the 12 parameter model (Barry and Hartigan 1987; Yang 1994), which the LogDet of Lockhart et al. (1987) and paralinear distances of Lake (1994) give an additive distance measure under; the 6 parameter model of Kimura (1981) with a consistent distance estimate given by Gojobori et al. (1982); and the 5 parameter model described in Takahata and Kimura (1981).

Discussion

Choosing a distance with which to estimate divergence times is not always straight forward. The only pairwise distance estimator which is linearly related to time under a stationary i.r. 12 parameter model is the LogDet (Steel 1994; Lockhart et al. 1994; Swofford et al. 1996) or paralinear distance (Lake 1994). This makes it is suitable for divergence time estimates using phylogenetic trees, given these conditions, plus a molecular clock and stationary base frequencies (Waddell 1995). Importantly though, if all sequences have a base composition close to equal frequency, then equation (1) may well return a distance just as additive (in expectation) as any under a stationary i.r. model, with the added advantage that equation (2) can be used to accommodate a variety of distributions of rates across sites. Of course, as base composition becomes unequal, but stationary, with site rates i.r. then the LogDet will become more useful for estimating relative divergence times. If, in addition, site rates are unequal the invariant sites-LogDet (Waddell 1995; see also Swofford et al. 1996; Swofford 1996) may become the best measure of relative divergence times using just pairwise distances. The problem with non-stationarity is that it becomes necessary to make a judgment of how severely the molecular clock is being violated and how much the LogDet distance is deviating from giving an unweighted estimate of the number of substitutions per site.

An additional use of the distance in equation (2) is for obtaining a first approximation to the length of edges on a tree when performing ML with rates across sites. This could be done by building a tree with a method such as weighted least squares (e.g. Fitch and Margoliash 1967) based on these pairwise distances (e.g. Adachi 1995, Swofford 1996). An alternative would be to use generalised parsimony (e.g. Sankoff 1975) to estimate the **P** matrix for each edge in the tree. Application of the transformation $-\mathrm{tr}(\Pi(M^{-1}[\mathbf{P}])$ to each edge could give a revised estimate of the length of that edge, better taking into account multiple hits, prior to the first iteration with likelihood. This is a more general application of a method being implemented in PAUP* (Swofford and Rogers, in preparation, Swofford 1996), where an important preparatory step is to assign parsimony changes to their most likely positions, i.e. on the longest edges.

All the methods suggested here can easily be extended to fewer or to more states e.g. purines versus pyrimidines (2 states), amino acids (20 states), the first 2 sites of protein codons (16 states, or 162 entries in **F**) or all 61 non-stop codons. The basic assumptions remain the same: the process is time reversible, or the base composition is equal-frequency, so all matrices have a double stochastic form. The use of 20 or 61 states removes much of the local correlations between site substitutions caused by the genetic code, and for this reason (plus the reduced likelihood of convergences) such distances may be preferable with more diverged sequences.

There is a tendency to accept that current models are adequate, for the simple reason we like to believe something is solid and certain in our modeling. While this may be largely justified in the case of very closely related species, figure 3a suggests there are problems extending our modeling back to older sequences. If so, it seems unlikely that any i.i.d. distance model based on nucleotides will be fully immune to such bias. This suggests that a shift to non-reversible or non-i.i.d. models may be necessary to infer information such as ti/tv ratios reliably from older sequences. Such biases are also expected to distort divergence time estimates. Since this is a major and growing use of phylogenetic trees, it is important that much more attention be paid to these matters.

Acknowledgments

Thanks to the Smithsonian Fellowships (1994) for paying travel and stipend for PJW to assist David Swofford to implement these methods into PAUP*. Thanks to David Penny for his comments, encouragement and support in completing the final manuscript.

References

Adachi J, Hasegawa M (1994) Time scale for the mitochondrial DNA tree of human evolution. In: Brenner S, Hanihara K (eds) The origin and past of modern humans as viewed from DNA. World Scientific, Singapore pp 46-68

Barry D, Hartigan JA (1987) Asynchronous distance between homologous DNA sequences. Biometrics 43:261-276

Chang JT (1996) Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. Math. Biosci. 134:189-215

Churchill GA, von Haeseler A, Navidi WC (1992) Sample size for phylogenetic inference. Mol. Biol. Evol. 9:753-769

Evans SN, Speed TP (1993) Invariants of some probability models used in phylogenetic inference. Annals of Statistics 21:355-377

Farris JS (1973) A probability model for inferring evolutionary trees. Syst. Zool. 22:250-256

Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401-410

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368-376

Felsenstein J (1982) Numerical methods for inferring evolutionary trees. Quart. Rev. Biol. 57:379-404

Felsenstein J (1984) Distance methods for inferring phylogenies: A justification. Evolution 38:16-24

Felsenstein J (1993) PHYLIP (Phylogeny Inference Package) and manual, version 3.5c. Department of Genetics, University of Washington, Seattle

Felsenstein J, Churchill GA (1996) A hidden Markov model approach to variation among sites in rate of evolution. Mol. Biol. Evol. 13:93-104

Fitch WM, Margoliash M (1967) Construction of phylogenetic trees. Science 155:279-284

Fitch WM, Markowitz E (1970) An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genetics 4: 579-593

Gaut BS, Lewis PO (1995) Success of maximum likelihood in the four-taxon case. Mol. Biol. Evol. 12:152–162

Gojobori T, Ishii K, Nei M (1982) Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. J. Mol. Evol. 18:414- 422

Golding GB (1983) Estimates of DNA and protein sequence divergence: An examination of some assumptions. Mol. Biol. Evol. 1:125-142

Gu X, Fu Y-X, Li W-H (1995) Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol. Biol. Evol. 12:546-557

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 21:160-174

Hasegawa M, Fujiwara M (1993) Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. Mol. Phyl. Evol. 2:1-5

Hendy MD, Penny D (1989) A framework for the quantitative study of evolutionary trees. Syst. Zool. 38:297-309

Hillis DM, Mable BK, Moritz C (1996) Applications of molecular systematics. In: Hillis DM, Moritz C, Mable BK (eds) Molecular Systematics, second edition. Sinauer Associates, Sunderland, Massachusetts, pp 515-543

Horai S, Satta Y, Hayasaka K, Kondo R, Inoue T, Ishida T, Hayashi S, Takahata N (1992) Man's place in the Hominoidea revealed by mitochondrial DNA genealogy. J. Mol. Evol. 35:32-43

Jin L, Nei M (1990) Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. 7:82-102

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian Protein Metabolism. Academic Press, New York, pp 21-132

Keilson J (1979) Markov chain models - rarity and exponentiality. Applied Mathematical Sciences, Vol. 28. Springer-Verlag, New York

Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111-120

Kimura M (1981) Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA 78:454-458

Lake JA (1994) Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. Proc. Natl. Acad. Sci. USA. 91:1455-1459

Lanave C, Preparata G, Saccone C, Serio G (1984) A new method for calculating evolutionary substitution rates. J. Mol. Evol. 20:86-93

Lewis PO, Swofford DL (1996) A general method for estimating evolutionary distances under any specific time reversible model. (in preparation)

Lockhart PJ, Larkum AW, Waddell PJ, Steel MA, Penny D (1996) Evolution of chlorophyll and bacteriochlorophyll: The problem of invariant sites in sequence analysis. Proc. Natl. Acad. Sci. USA 93 (in press)

Lockhart PJ, Steel MA, Hendy MD, Penny D (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11:605-612

Olsen GJ (1987) The earliest phylogenetic branchings: Comparing rRNA-based evolutionary trees inferred with various techniques. Cold Spring Harbor Symp. Quant. Biol. 52:825-837

Read TRC., Cressie NAC (1988) Goodness-of-fit statistics for discrete multivariate data. Springer-Verlag, New York

Reeves JH (1992) Heterogeneity in the substitution process of amino acid sites of proteins coded for by Mitochondrial DNA. J. Mol. Evol. 35:17-31

Ritland K, Clegg MT (1987) Evolutionary analysis of plant DNA sequences. The American Naturalist 130, supplement:S74-S100

Rodríguez F, Oliver JL, Marin A, Medina JR (1990) The general stochastic model of nucleotide substitution. J. Theor. Biol. 142:485-501

Saitou N (1990) Maximum likelihood methods. In: Doolittle RF (ed) Methods in enzymology. vol 183. Molecular evolution: computer analysis of protein and nucleic acid sequences. Academic press, San Diego. pp 584-598

Sankoff D (1975) Minimal mutation trees of sequences. SIAM J. Appl. Math. 28:35-42.

Shoemaker JS, Fitch WM (1989) Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. Mol. Biol. Evol. 6:270-289

Sidow A, Nguyen T, Speed TP (1992) Estimating the fraction of invariable codons with a capture-recapture method. J. Mol. Evol. 35:253-260

Steel MA (1994) Recovering a tree from the leaf colourations it generates under a Markov model. Appl. Math. Lett. 7:19-23

Steel MA, Székely L, Erdös PL, Waddell PJ (1993) A complete family of phylogenetic invariants for any number of taxa under Kimura's 3ST model. New Zealand Journal of Botany (Conference Issue) 31:289-296

Stuart A, Ord JK (1987) Kendall's advanced theory of statistics. Volume 1. 5th ed. Edward Arnold, London

Stuart A, Ord JK (1991) Kendall's advanced theory of statistics. Volume 2, Distribution theory. Classical inference and relationship. 5th ed. Edward Arnold, London

Swofford DL (1996) *PAUP** version 4.0. Sinauer Associates, Sunderland, Massachusetts (in press)

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM (1996) Phylogenetic Inference. In: Hillis DM, Moritz C, Mable BK (eds) Molecular Systematics, second edition. Sinauer Associates, Sunderland, Massachusetts, pp 407-514

Takahata N, Kimura M (1981) A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. Genetics 98:641-657

Tamura K (1992) Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol. Biol. Evol. 9:678-687

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10:512-526

Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures on Mathematics in the Life Sciences 17:57-86

Waddell PJ (1995) Statistical methods of phylogenetic analysis: Including Hadamard conjugations, LogDet transforms, and maximum likelihood. Ph.D. thesis, Massey University

Waddell PJ, Penny D (1996) Evolutionary trees of apes and humans from DNA sequences. *In* Handbook of Human Symbolic Evolution. (ed. A.J. Lock and C.R. Peters) Clarendon Press, Oxford. (in press since August 1993, preprints available)

Waddell PJ, Penny D, Moore T (1996) Extending Hadamard conjugations to model sequence evolution with variable rates across sites. Preprint Series of Dept. of Mathematics, Massey University

Yang Z (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10:1396-1401

Yang Z (1994) Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39:105–111

Zharkikh A (1994) Estimation of evolutionary distances between nucleotide sequences. J. Mol. Evol. 39:315-329

Appendix 1

(a) Under a time reversible model, with stationary root distribution, **F** is always symmetric.

Proof:

If **R** is reversible, $\mathbf{F} = exp[\mathbf{R}\tau_1]^t \Pi exp[\mathbf{R}\tau_2]$

$= \Pi exp[\mathbf{R}(\tau_1 + \tau_2)]$

$\mathbf{F}^t = exp[\mathbf{R}\tau_2]^t \Pi^t exp[\mathbf{R}\tau_1]$

$= \Pi\, exp[\mathbf{R}\tau_2]\, exp[\mathbf{R}\tau_1]$

$= \Pi \exp[\mathbf{R}\tau_1 + \tau_2]$

so, $\mathbf{F} = \mathbf{F}^t$.

(b) Proof that symmetrising $\hat{F}$ gives the ML estimate of **F** as $\mathbf{F}^{\#}$.

Since $c\mathbf{F}$ is expected to be symmetric under the model, then $cF_{ij} = cF_{ji}$. As $c\mathbf{F}$ is expected to have a multinomial distribution (giving binomial marginal distributions for each entry cF_{ij}), the ML estimator of $cF_{ij} + cF_{ji}$ is $c(\hat{F}_{ij} + \hat{F}_{ji})$, so it follows that the ML estimator of F_{ij} or F_{ji} is 1/2 $(\hat{F}_{ij} + \hat{F}_{ji})$ i.e. entry $F^{\#}_{ij}$. This applies jointly for all entries in $\mathbf{F}^{\#}$ since all symmetrised pairs are non-overlapping.

(c) Proof that for any Markov process operating on a rooted tree obeying a molecular clock, then **F** is symmetric (without assuming that the root base composition is necessarily in equilibrium).

Under a molecular clock,

$\mathbf{F} = \mathbf{P}^t\Pi\mathbf{P}$, and so

$\mathbf{F}^t = \mathbf{P}^t\Pi(\mathbf{P}^t)^t$

$= \mathbf{F}$, as claimed.

Appendix 2

Proof of equation (2).

We have, $\mathbf{F} = E\lambda[\Pi exp[\mathbf{R}\tau\lambda]] = \Pi E\lambda[exp[\mathbf{R}\tau\lambda]] = \Pi M[\mathbf{R}\tau]$,

by the relationship $\mathbf{P} = M[\mathbf{R}\tau]$ of equation 4. Thus, $\mathbf{R}\tau = M^{-1}[\Pi^{-1}\mathbf{F}]$, and since

$\delta_{ij} = -\text{tr}(\Pi\mathbf{R})\tau E\lambda[\lambda]$, and we are assuming that $E\lambda[\lambda]=1$, we have $\delta_{ij} = -\text{tr}(\Pi M^{-1}[\Pi^{-1}\mathbf{F}])$,

as claimed.

Appendix 3

Proof that correcting sequence dissimilarity under the identical rates (i.r.) assumption will (for long sequences) underestimate the divergence δ_{ij} between two sequences *i* and *j* if the rates vary across sites.

Let $d_{ij}(k)$ denote the probability that sequences i and j have a different state at site k. So, if d_{ij} is the expected dissimilarity between the two sequences (the expected proportion of sites where the two sequences differ) we have $d_{ij} = \frac{1}{c}\sum_{k} d_{ij}(k)$, where c is the length of the two sequences. Now, from standard results concerning reversible Markov processes (Keilson, 1979) it is easily shown that

$$d_{ij}(k) = H[\lambda_k \tau_{ij}]$$

where H is a function of the form:

$$H[x] = \alpha_0(1 - \sum_{t=1}^{r-1} a_t exp[-\mu_t x])$$

in which the constants α_t and μ_t are non-negative, and dependent only on the form of the $r \times r$ rate matrix $\mathbf{R}$ (for example, in the Jukes-Cantor model, or Felsenstein-Tajima-Nei model, where $H[x] = b(1 - exp[-x / b])$, where $b = 1- (\pi_A^2 + \pi_C^2 + \pi_G^2 + \pi_T^2)$(see Swofford et al. 1996).

In particular, H is a monotone increasing function, and so possesses a left inverse, H^{-1}. Now, if all the sites evolve with rate $\lambda_k = 1$, then $\delta_{ij} = -tr(\Pi\mathbf{R})\tau_{ij} = -tr(\Pi\mathbf{R})H^{-1}[d_{ij}]$. On the other hand, if the distribution of the rates is nondegenerate, but with the same mean (viz. 1) then

$$H^{-1}[d_{ij}] = H^{-1}[\frac{1}{c}\sum_{k} d_{ij}(k)\,] < \frac{1}{c}\sum_{k} H^{-1}[d_{ij}(k)] = \frac{1}{c}\sum_{k}\lambda_k\tau_{ij} \quad = \tau_{ij}$$

where the inequality arises due to Jensen's inequality which applies since H^{-1} is strictly convex. Thus, $-tr(\Pi\mathbf{R})H^{-1}[d_{ij}]$ underestimates δ_{ij} whenever the distribution is nondegenerate, (while $-tr(\Pi\mathbf{R})H^{-1}[d_{ij}] = \delta_{ij}$ if the sites evolve at the same rate), as claimed. Thus, this proof generalises the result of Golding (1983) beyond the Kimura 2ST model.

Appendix 4.

Proof of equation (4).

First we recall how the domain of the moment generating function M is extended so as to be defined on matrices. Namely, if $M[x] = 1 + \sum_{k=1}^{\infty}\lambda_k x^k$ then for a matrix $\mathbf{X}$,

$$M[\mathbf{X}] := \mathbf{I} + \sum_{k=1}^{\infty}\lambda_k \mathbf{X}^k .$$

We may assume, without loss of generality, that the rate matrix $\mathbf{R}$ is diagonalizable, so that we can write

$$\mathbf{R}\tau = \mathbf{A}\mathbf{D}\mathbf{A}^{-1}$$

Then $\mathbf{P} = E_\lambda[exp(\mathbf{R}\tau\lambda)] = E_\lambda[\mathbf{A}exp(\lambda\mathbf{D})\mathbf{A}^{-1}] = \mathbf{A}E_\lambda[exp(\lambda\mathbf{D})]\mathbf{A}^{-1}$

Now, $exp(\lambda\mathbf{D})$ is the diagonal matrix with ii-th entry $exp(\lambda D_{ii})$, thus $E_\lambda[exp(\lambda\mathbf{D})] = M(\mathbf{D})$, by the above definition of $M(D)$. Thus, again invoking this definition:

$$\mathbf{P} = \mathbf{A}\, M[\mathbf{D}]\mathbf{A}^{-1} = M[\mathbf{A}\mathbf{D}\mathbf{A}^{-1}] = M[\mathbf{R}\tau],$$

as claimed.

Appendix 5

Proof of the delta method approximation, equation (5), for the variance of equation (2).

The proof is a direct extension of Barry and Hartigan's (1987) proof of the special case where $M[x] = e^x$ to a general M. In particular, by equation 2 and 4, we have

$$\delta_{ij} = \mathrm{tr}(\Pi M^{-1}[\mathbf{P}]) = -\sum_{r=1}^{\infty} a_r \mathrm{tr}(\Pi \mathbf{B}^r)$$

and the remainder of Barry and Hartigan's proof applies upon substitution of their term 1/r for a_r.

Appendix 6

Example of the derivation of the coefficient for use in the estimation of variance, using the inverse Gaussian distribution of rates

The term a_r is given by the equation $M^{-1}[1\text{-}x] = -\sum_i a_i x^i$. For the inverse Gaussian distribution, $M^{-1}[1\text{-}x]$ can be written as $y = \frac{d}{2}\left[1-\left\{1-\frac{\ln[1-x]}{d}\right\}^2\right]$. The function

$$ln[1\text{-}x] = -\sum_{i=1}^{\infty}\frac{x^i}{i} \text{ so } y = \frac{d}{2}\left[1-\left\{1+\frac{1}{d}\sum\frac{x^i}{i}\right\}^2\right] = -\sum\frac{x^i}{i}-\frac{1}{2d}\left(\sum\frac{x^i}{i}\right)^2.$$

Thus, $a_i = \frac{1}{i}+2d\sum_{j=1}^{i-1}\frac{1}{j(i-j)}$

Appendix 7

Proof of the test of time reversibility via equalities E1-E3 (marked as equations (6) to (8)), and that **R** can be written as $\Pi\mathbf{S}$ under a time reversible model.

From Tavaré (1986), Barry and Hartigan (1987), or Rodríguez et al. (1990) we have that **R** forms a reversible model precisely if

$$\mathbf{R}^t\Pi = \Pi\mathbf{R}, \qquad (1)$$

where Π is a diagonal matrix with elements $(\pi_1, \pi_2, \pi_3, \pi_4)$, and where π is the equilibrium vector for **R** - that is $\pi\mathbf{R} = 0$, and the components of π are non-negative and sum to 1.

Now, set $\mathbf{S} = \Pi\mathbf{R}$,

then $\mathbf{S} = \mathbf{S}^t$ by (1) so that we can write,

$$\mathbf{S} = \begin{bmatrix} * & a & b & c \\ a & * & d & e \\ b & d & * & f \\ c & e & f & * \end{bmatrix},$$

(where the row (and column) sums are zero (so we again omit writing out each diagonal entry, * represents minus sum of the off diagonal elements in a row).

Now, since $\mathbf{R} = \Pi^{-1}\mathbf{S}$ we have,

$$\mathbf{R} = \begin{bmatrix} * & x_1a & x_1b & x_1c \\ x_2a & * & x_2d & x_2e \\ x_3b & x_3d & * & x_3f \\ x_4c & x_4e & x_4f & * \end{bmatrix} \qquad (2)$$

where $x_i = \pi_i^{-1} \neq 0$.

Conversely, any rate matrix which can be written in the form (2) with $x_i \neq 0$ for i = 1, .., 4 forms a reversible model, since we may assume the x_i's and the values a-f are positive, and if we let $\mu = \left(\frac{1}{x_1}, \frac{1}{x_2}, \frac{1}{x_3}, \frac{1}{x_4}\right)^{-1}$, then $\pi = \left(\frac{\mu}{x_1}, \frac{\mu}{x_2}, \frac{\mu}{x_3}, \frac{\mu}{x_4}\right)^{-1}$ is the equilibrium vector for **R** and $\mathbf{R}^t\Pi = \Pi\mathbf{R}$ as required.

Thus the matrix $\begin{bmatrix} * & A & B & C \\ D & * & E & F \\ G & H & * & I \\ J & K & L & * \end{bmatrix}$ is reversible if and only if it can be written in the form (2) for x_i's$\neq 0$. But then $AGE = (x_1a)\,(x_3b)\,(x_1d) = x_1\,x_2\,x_3abd$,

and $BDH = (x_1b)\,(x_2a)\,(x_3d) = x_1\,x_2\,x_3abd$,

that is AGE = BDH, while similarly AJF = CDK and EKI = FHL (E1 - E3).

Conversely, suppose the three equations hold. We show that **R** can be written in the form (2) and therefore it forms a reversible model.

Set $x_1 = \frac{C}{J}, x_2 = \frac{EI}{HL}, x_3 = \frac{I}{L}, x_4 = I,$ and

$a = \frac{AJ}{C}, b = \frac{BJ}{C}, c = J, d = \frac{HL}{I}, e = \frac{FHL}{EI}, f = L.$

Then clearly A = x_1a, B= x_1b, C = x_1c, H = x_3d, I = x_3f, J = x_4c, L = x_4f, E = x_2d, F = x_2e.

It remains to check that,

D = x_2a,

G = x_3b,

K = x_4e,

We have,

$$x_2a = \frac{EI}{HL}\frac{AJ}{C} \overset{(E3)}{=} \frac{FAJ}{KC} \overset{(E2)}{=} \frac{CDK}{KC} = D,$$

$$x_3b = \frac{I}{L}\frac{BJ}{C} \overset{(E1)}{=} \frac{AGEIJ}{DHLC} \overset{(E2)}{=} \frac{GEIK}{HLF} \overset{(E3)}{=} \frac{GFHL}{HLF} = G,$$

$$x_4e = \frac{FHL}{EI} \overset{(E3)}{=} \frac{EKI}{EI} = K,$$

as required, completing the proof.

Note:

If $\mathbf{R}$ is reversible, then we can write $\mathbf{R} = \mathbf{Q}\Pi$ (for some symmetric $\mathbf{Q}$) in place of $\mathbf{R} = \Pi^{-1}\mathbf{S}$.

Proof: Set $\mathbf{Q} = \mathbf{R}\Pi^{-1}$. We need to show $\mathbf{Q} = \mathbf{Q}^t$. Since $\Pi\mathbf{R} = \mathbf{R}^t\Pi$, if we pre and post multiply this equation through by Π^{-1} we get, $\mathbf{Q} = \mathbf{R}\Pi^{-1} = \Pi^{-1}\mathbf{R}^t = \mathbf{Q}^t$. Thus $\mathbf{Q} = \mathbf{Q}^t$ as claimed. Further, this shows $\mathbf{R}$ is reversible if and only if $\mathbf{R} = \mathbf{Q}\Pi$ (as Zharkikh 1994 notes).